Enstore and Serial Media

What is Enstore?

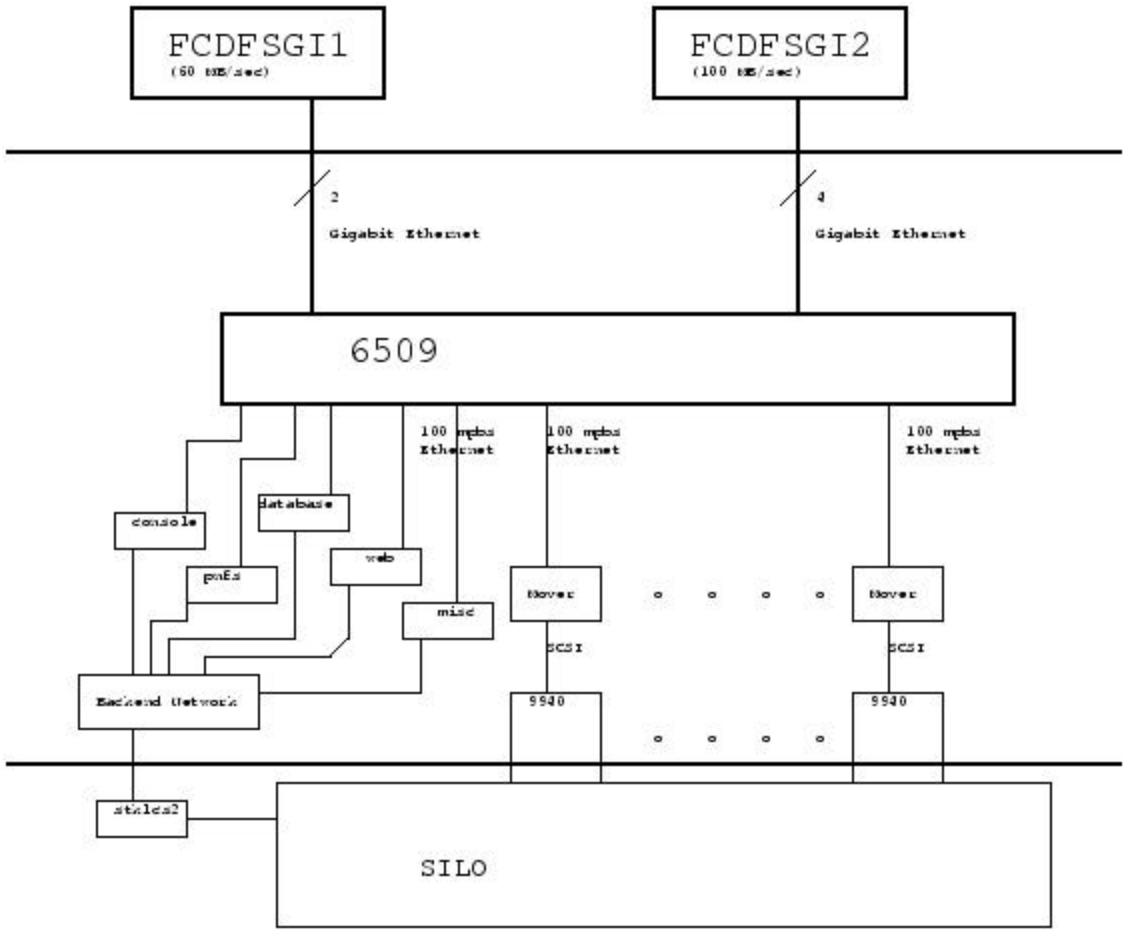
- Distributed software to move files to and from tape over the network.
- Two levels of access:
  - encp -- used within the FNAL site, with deterministic, tape-rate performance.
  - dCache -- “buffered”, off-site, and low performance onsite access.
- Usage: 12 storage groups ~= “experiments”

How Much Experince? Amount of Data Written

Technology	**Volume Today**	**Addl Tested**
STK 9940	15.9 TB	30 TB
STK 9840	24.3 TB	
LTO	3.3 TB	30 TB
M2	36.0 TB	40 TB
M1	12.2 TB	
AIT2	1.3 TB	20 TB
DLT	(epsilon)	
TOTAL	**93.0 TB**	**80 TB**

What is the data transfer interface?

- Encp:
  - File catalog is transported using NFS2 as a transport (pnfs from DESY).
    - ls -alt /pnfs/test/cdf/EG0001.2
  - Files are actually read and written using a command, “encp,” modeled after the unix “cp” command.
    - encp /pnfs/test/cdf/EG0001.2/ar01f36d.00186nx1 /dev/null
    - File Data are never moved over NFS
- FTP protocols for remote and buffered users.
  - Python FTP Client available (kftppy)
  - Kerberos FTP Kit, of any which will transport credentials using AUTH GSSAPI
  - Weak FTP, with restrictions, at option of the experiment.



Enstore & Serial Media

File Data are moved over a LAN

- Run II aggregate tape bandwidths are tiny compared to bandwidth of installed Cisco 6509 Switch.
- Enstore can transfer to/from performant linux farms directly.
- Enstore can feed big SGI's scalably.
  - 30 MB/sec per dedicated CPU.
- Attention must be paid to data movement LAN.

Usability Philosophy

- Error handling is a significant fraction of data movement software.
- Enstore has code:
  - to detect faults and alarm
  - to disclose its state to both administrators and experimenters.
- This represents a significant amount of work done in the package.

Fault Handling in Enstore:

- Broken components are detected…
  - Drives having many errors configure themselves out of the system.
  - Tapes having errors on many drives are flagged for investigation.
- Internal Monitoring.
  - Time series (e.g. Bakken plots)
    - http://www-stken.fnal.gov/enstore/cron_pics.html
- Alarms -- List of work for normal administration.
    - E.G. http://www-stken.fnal.gov/enstore/enstore_alarms.html
  - Faults -- Integrated into 24x7 pager system.
    - http://www-stken.fnal.gov/enstore/enstore_saag.html
- Automated drive cleaning.
  - Cleaning bit and other policies

Exposure of Internals to Experimenters

- http://www-stken/enstore/
  - Overall status -- up or down? Why?
  - Encp history -- recent transfers.
  - List currently working drives.
  - Lists of any file transfers waiting for drives.
- Shell tools to access each server.
- Animation program for internal administrations, and “heros”

Grid and dCache

- dCache -- disk based cache in front of Enstore.
  - Enables offsite access.
  - Built collaborative with DESY.
  - Optional if you want it.
- Built on Scalable pools of disk on independent computers.
- In early production now, but development is unfinished.
  - Read is the main emphasis.
  - Writes to tape are “crippled.”
  - FTP (w/ GSSAPI) for kerberos authenticated R/W access.
  - FTP (weak authentication) for read access only after Jan 1.
- Working on Grid access protocols
  - w/ANL over GridFTP protocol.
  - W/JLab, LBL on HRM protocol.
  - Grid work is constant work, protocols are not stable.

Hardware Strategy

- Lab strategy:
  - Flexible choice of best tape technology.
  - Automate tape handling.
- Enstore support:
  - Multiple Libraries.
    - Support of STK and AML libraries allows many drives.
  - Multiple drives.
    - Vanilla use of command set.
  - File access abstraction.
    - Allows experiments to change tape technology.
    - Allows migration media as time passes.

Current Hardware Direction 9940

- Characteristics:
  - Automated in STK powderhorn @ $18/slot ($85K/ > 5000 slots)
  - Media is $79/ 60 GB cartridge, physically robust.
  - Drive is $27K, 10 MB/sec read and write.
- Experience:
  - In production at FNAL: **No Known Defects**.
  - HEP labs pleased with production experience.
  - Very good experience when drives were in early part of life cycle.
- Upgrade:
  - Q3 2002: 200 GB/cart, 30 MB/sec.
  - MUST re-buy drives
  - Uses identical media

Current Tape Direction - LTO

- Characteristics
  - Chance for multi vendor drives, media.
  - Automated in ADIC AML/2 @ $66/slot.
  - Media is ~$120/100 GB, DLT like cartridge.
  - IBM Drive was ~$8K, 15 MB/sec Read and Write, installed in AML/2.
- Experience
  - Very little production experience, slated for D0 Monte Carlo.
  - 30 TB evaluation test uncovered problems:
    - Drive crash on writes.
    - Can see the BER on reads ~ 10 TB (we catch this).
- Upgrade
  - to 200 GB/cart, different cartridge, higher speed.

Deprecated Hardware: Mammoth

- Media Experience:
  - The mechanical (and unchangeable per ADIC) limits in the AML/2 do not protect 8mm cartridges. There is risk of automated systematic damage.
    - Lost alignment in AML/2 has smashed cartridges into fixtures..
      - Fragile cassettes have been damaged and can be damaged en mass.
    - Duplicate VSN? AML/2 will file it in an already occupied slot.
    - Insert an inverted cartridge? AML/2 will insert in the drive upside down.
  - Mis-inventories of library are associated with narrowness of cartridge.
    - We trust the pick of a blank to be accurate.
  - Have seen recalled ME media.
    - Lot control numbers occluded by bar-code label needed for automation.
  - Media chafing, shutter damage due to mis teaching of the AML
- This would be true of all 8mm automated in an AML/2.
  - Cannot write code to detect/prevent this type of damage in all cases.

Evaluated Hardware: AIT-2

- Burn-in:
  - Made a tape which we could not read in any of 6 drives
  - The cleaning bit is never set.
    - The drives performed very badly when not cleaned intensively.
    - The manual says “don’t clean”. Excessive cleaning normally affects lifetime.
- 30 TB transfer test:
  - “Pathological cleaning” necessary.
    - Waiting for cleaning bit proved tp be disasterous
  - Recorded results excludes errors that enstore nominally re-tries on.
  - Lost data due to tape positioning error. (33 files, one incident)
    - Spacing file marks is intrinsic to enstore’s design.
- Static:
  - **All the defects of any 8mm AML/2 integration**.
  - Multiple drives/ bus - No shelter from SCSI bus resets.
  - Cannot access diagnostic port to work errors.

Can this benefit CDF?

- It is likely enstore can be grafted onto CDF DHS.
  - Run II scale experiments were considered in its design.
  - CDF architecture must change to exploit enstore.
  - This change is not necessarily bad, and could likely be substantially incremental.
- Good operational and administrative process exist within enstore.
  - Having left 8mm behind, there is adequate staffing in ISD for a CDF STK tape plant.
- Grid attachment and other future oriented work has begun.
  - Will bring data to experiment's institutions.
- Good hardware configurations are known, and are essential for adequate service to an experiment.
- ISD is unconvinced and very skeptical of the operational properties of the AIT-2/AML-2 tape plant, we can give no particular assurances to CDF for usability, schedule.
  - ISD feels an STK plant will give the experiment very good service.
- ISD feel that it is likely that running the current DH system on better drives would improve its operational properties.